

Stabilization of infrared image sequence with rotation, scaling and view angle changes

Daniel McReynolds and Yunlong Sheng
COPL, Laval University, Ste-Foy, Qc, G1K 7P4, Canada

Langis Gagnon
Lockheed Martin Electronic Systems Canada, 6111 Royalmount Ave., Montreal, Qc, H4P 1K6, Canada

Léandre Sévigny
Defence Research Establishment Valcartier, 2459 Boul. Pie XI nord, C.P. 8800, Courcellette, Qc, G0A 1R0, Canada

ABSTRACT

An enhanced greylevel differential invariant matching scheme is applied to the stabilization of real-world, infrared image sequences that have large translation, rotation, scaling and viewpoint changes. Its performance is compared with that of Zhang's robust image matching method.

Keywords: Image stabilization, differential invariants, epipolar geometry, infrared, scale-space, k-d tree

1. INTRODUCTION

Image stabilization is the image registration applied to one video image sequence from a single camera. When the camera is mounted on an unsteady or moving platform and objects are far from the camera, the 3-D space motion of the camera will affect the images. Even small movements of the platform can result in large displacements of the images. This effect should be estimated and eliminated electronically by warping each image frame into precise alignment with a reference frame. The displacement between images has a predominantly global character. In most cases the distortion between two consecutive frames is a 2-D rigid body translation and rotation. The third dimension movement changes the distance between camera and object, resulting in a change of scale in the consecutive images. There is also motion of the target with respect to the stationary scene background. This motion is, however, useful and is not to be corrected by the stabilization process.

Theoretically, the motion estimation from a time sequence of monocular images with a single view is an ill-posed problem because the solution is not unique. The fact that the camera has a limited aperture also results in the non-uniqueness of the solution. Until now, the primary means available to stabilize images from a camera on a moving vehicle has been to mount the camera on an electro-mechanical stabilizing platform. These stabilizers are bulky and expensive. Their performance degrades with vibration in the critical 0 – 20 Hz range. However, the introduction of real-time electronic image stabilization systems into the field is becoming prevalent as evidenced by a recent special issue of *Real-Time Imaging* dedicated to this topic (Real-Time Imaging 1996, Maheux 1998). An automatic electronic image stabilization system should first estimate components of scene motion that are due to camera motion, and then eliminate these components. A temporal filter (frame difference) then eliminates the background scene while highlighting targets that are tracked over multiple frames.

Image registration methods can be area-based or feature-based, also referred to as block matching or attribute matching. Block matching uses cross-correlation. The full image information is utilized. The method can be applied to any type of image, rich or poor in texture. The block correlation methods are robust against random noise and have high accuracy. However, block matching is expensive in computation time. The computational cost becomes prohibitive when the image displacement is large. The main advantages of the feature-based methods are their speed and ability to account for rotation, shearing and other image deformations. However, the feature-based methods will fail to find matches in structure-less areas. The reliability of those methods depends on the reliability of feature extraction process.

In this paper we report on an experimental investigation of two state-of-the-art computer vision point-based image matching techniques applied to the problem of image stabilization on real-world infrared image sequences – images which are generally noisy with low contrast. The frame to frame motion can be very large with significant scale change, rotation and projective distortion due to the camera motion. To address the issue of scale and rotational invariance, the method of greylevel differential invariant matching (GDI) is tested (Schmid and Mohr 1995). Two extensions to GDI (McReynolds 1997) are experimentally validated. Matching speed is increased by searching over a space of differential invariant vectors with k -d trees so that a query finishes in logarithmic expected time. Scale-space tracking of matches is utilized which significantly improves the ratio of true to false matches.

An alternative method due to Zhang *et al.* is based on geometric verification with the epipolar constraint (Zhang, Deriche et al 1995). However, the initial matching is not scale and rotation invariant. The performance of the two methods

is experimentally investigated and suggestions for future work are given. In particular, we note that the GDI matching is sensitive to projective distortion and the method of Zhang *et al.* requires a better technique for finding the initial matches before epipolar geometry verification.

Experimental assumptions include 1) pixels are square, 2) image overlap is greater than 25 percent, and 3) intrinsic camera parameters are unknown. Experiments with both algorithms produce good results for the most image sequences. The evaluation metric is the percentage of correct matches of the total number of matches found. The mismatch rate is approximately 20 percent for the image pair reported here due to the large change in viewpoint. Matching images earlier in the sequence to the same reference yields a mismatch rate of 5 to 15 percent. This robustness is important for registering real world images. The resultant mismatches in the GDI method are easily handed in the following step by global transformation fitting or an epipolar verification step such as the one incorporated in the method due to Zhang *et al.*

2. TWO IMAGE MATCHING METHODS

The greylevel differential invariant matching method is feature point based. It is assumed that the images are formed by a perspective projection. The method of Zhang *et al.* is a mixture of feature and area based methods. For both algorithms, the feature points are first extracted with the Harris-Stephens corner detector that is based on finding local maxima in the Gaussian curvature of the local autocorrelation of the brightness function (Harris and Stephens 1988). For image matching the two methods are not mutually exclusive either. The GDI hypothesized matches may be verified using the epipolar constraint. The method of Zhang *et al.* is one such method for the uncalibrated perspective camera, and the method of McReynolds and Lowe (McReynolds and Lowe 1996) is another for the calibrated perspective camera.

2.1 Greylevel differential invariant matching

At each feature point in the reference and current image found by the Harris-Stephens corner detector, a GDI vector is computed. Most feature based image matching methods assume a rigid transformation and a diffuse local surface reflectance. The GDI representation and matching method (Schmid and Mohr 1995) are invariant to image rotation, scaling and translation. A normalized version of the representation is also invariant to brightness scaling. For the perspective projection of a locally rigid 3D transformation, the local 2D projected motion at a non-boundary point \mathbf{x}_0 can be modeled by a rigid transformation if the rotation in depth and the depth range in a small neighbourhood of \mathbf{x}_0 are not large relative to the depth of \mathbf{x}_0 (Weng *et al.* 1992). This analysis lays the foundation for the use of local differential invariants as motion invariant point attributes for characterizing the local brightness distribution at a feature point. The representation is fairly robust with respect to rotation in depth that leads to foreshortening of surface patches, i.e., in general, a local affine distortion of the brightness surface.

The local jet, \mathbf{J} , of order N is defined at a point $\mathbf{x} = (x_1, x_2, \dots, x_D)^T$ ($D = 2$ for an image) as

$$\mathbf{J}^N [\mathbf{I}](\mathbf{x}, \sigma) = \{L_{i_1, \dots, i_n}(\mathbf{x}, \sigma) \mid (\mathbf{x}, \sigma) \in \mathbf{I} \times \mathfrak{R}^+; n = 0, \dots, N\}$$

where \mathbf{I} is the image array and σ is a given scale. $L_{i_1, \dots, i_n}(\mathbf{x}, \sigma)$ is the convolution of image \mathbf{I} with the Gaussian derivatives $G_{i_1, \dots, i_n}(\mathbf{x}, \sigma)$ with

$$G(\mathbf{x}, \sigma) = \frac{1}{2\pi \sigma^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}} \text{ and } G_{i_1, \dots, i_n}(\mathbf{x}, \sigma) = \frac{\partial^n}{\partial i_1 \dots \partial i_n} G(\mathbf{x}, \sigma)$$

where $n = 0, 1, 2, \dots$ and $i_k \in \{x_1, x_2, \dots, x_D\}$ for $k = 1, \dots, n$.

The local jet represents the truncated Taylor series expansion of the image function and is useful for encoding the position dependent geometry of the brightness surface. The Taylor series expansion requires the computation of the partial derivatives which is an ill-posed problem for noisy images in the sense of Hadamard, since differentiation is discontinuous in the presence of noise. The calculation of the derivative requires the specification of an operator. The choice of the derivative of the Gaussian is motivated by the properties of the Gaussian kernel. It is smooth, has the smallest product of localization uncertainty in space and frequency, and, it is the least likely operator to introduce artifacts in the filtered image (Marr 1982, Horn 1986). As an operator, it can be applied efficiently in the sense that it is the unique rotationally symmetric smoothing operator that is separable into the product of two one-dimensional operators (Horn 1986). The Gaussian function has the property of scale dependency that is given by σ .

Adopting the notation of Schmid and Mohr (Schmid and Mohr 1995), the differential invariant vector V , is given by

$$V = \begin{bmatrix} L \\ L_x L_x + L_y L_y \\ L_{xx} L_x L_x + 2L_{xy} L_x L_y + L_{yy} L_y L_y \\ L_{xx} + L_{yy} \\ L_{xx} L_{xx} + 2L_{xy} L_{xy} + L_{yy} L_{yy} \\ L_{xxx} L_y L_y L_y + 3L_{xyy} L_x L_x L_y - 3L_{xxy} L_x L_y L_y - L_{yyy} L_x L_x L_x \\ L_{xxx} L_x L_y L_y + L_{xxy} (-2L_x L_x L_y + L_y L_y L_y) + L_{xyy} (-2L_x L_y L_y + L_x L_x L_x) + L_{yyy} L_x L_x L_y \\ L_{xxy} (-L_x L_x L_x + 2L_x L_y L_y) + L_{xyy} (-2L_x L_x L_y + L_y L_y L_y) - L_{yyy} L_x L_y L_y + L_{xxx} L_x L_x L_y \\ L_{xxx} L_x L_x L_x + 3L_{xxy} L_x L_x L_y + 3L_{xyy} L_x L_y L_y + L_{yyy} L_y L_y L_y \end{bmatrix}$$

where $i_1, \dots, i_n \in \{x, y\}$ on which the image brightness function, \mathbf{I} , is defined. For example, L is the average brightness, $L_x L_x + L_y L_y$ is the gradient magnitude squared, and $L_{xx} + L_{yy}$ is the Laplacian of the brightness function. The components of the vector, V , of invariants, are the complete and irreducible set of differential invariants up to third order. These functions are rotationally symmetric.

Differential invariants can be also invariant to an affine transformation of the brightness function given by $\tilde{\mathbf{I}}(x, y) = a\mathbf{I}(x, y) + b$. These invariants are the last seven components of the differential invariant vector, V , normalized by an appropriate power of the gradient magnitude squared, i.e., $(L_x L_x + L_y L_y)^p$, where the power p is determined so that the ratio $a^k/a^{(2p)}$ equals one, where k is the power of the scaling coefficient a for a particular product of the L_{i_1, \dots, i_n} . For example, the first component of the brightness affine invariant vector is given by

$$\frac{L_{xx} L_x L_x + 2L_{xy} L_x L_y + L_{yy} L_y L_y}{(L_x L_x + L_y L_y)^{3/2}}$$

where the scaling coefficient of the numerator due to a scaling of the brightness function by a can be shown to be a^3 . Note that the differential invariants, other than L , by dint of differentiation are invariant to the brightness translation b .

The multi-scale representation consists of a set of differential invariant vectors computed over a range of scales centered on a base scale σ_0 that vary by a factor of $1.2n$ for some integer range of n . The 20 percent factor is empirically derived and reflects the expected differential scale range over which the invariants do not change appreciably. The complete set of scales is given by $\sigma_i = (6/5)^i \sigma_0$ where $i \in (-n \dots -1, 0, 1 \dots n)$. A value of four for n yields the scale factor range 0.48 to 2.07, hence there are nine differential invariant vectors for each keypoint.

For each point in each image the differential invariant vector at the reference scale σ_0 is used as the query vector for matching the differential invariant vectors in the other images. The Mahalanobis distance is used to determine the nearest neighbour. Points are declared to be corresponding when the pair of points are mutually selected as closest, i.e., they are the closest match to each other from image one to two and vice versa. The space of differential invariant vectors can be organized in a hash table (Schmid and Mohr 1995) or with a tree representation such as the k -d tree. Nearest neighbour searching with k -d trees reduces the search time to logarithmic expected time. Sproull (Sproull 1991) gives the expected number of records examined as $R(k, b) \approx b\{[G(k)/b]^{1/k} + 1\}^k$, where k is the record dimension, b is the number of records in a bucket, and $G(k) \geq 1$ is a constant that accounts for the geometric properties of the norm used for the distance measure. Sproull assumes $G(k) = 1$. For our implementation of GDI matching, k is 7 and b is 1, hence $R(7, 1) = 128$.

A local multiscale analysis is used to filter out scale-space unstable and hence potentially incorrect matches. Differential invariants are computed at three reference scales that differ by multiples of ten percent of the reference base scale, i.e., $\sigma_{base}(1 + k * 0.1)$ for $k=0, \dots, 2$. The value of ten percent is half the expected scale sensitivity of differential invariants. Experimental results indicate that nearest neighbour matching with differential invariants is scale sensitive over moderate viewpoint changes. This value was found experimentally to yield good results. Too large a value eliminates most matches and too small a value does not effectively eliminate scale unstable matches.

The GDI method produces a set of point feature matches possibly including mismatches. The mismatches are automatically edited out using a relaxation process that eliminates those matches whose motion is not consistent with its

nearest neighbours for a given error bound. The resulting set of matches is used to perform a least squares estimate of the global affine transformation between the reference image and the current image.

With the affine parameters thus estimated the current image is interpolated at the location that corresponds to each pixel in the reference image. For the results given below bilinear interpolation is used. This interpolation process is the most time consuming part of the method and, in order to speed up the computation, the images are smoothed and sub-sampled versions of the original images by a factor of 4 yielding a dimension of approximately 128x128 pixels.

2.2 Epipolar constraint image matching for uncalibrated camera

This approach combines classical normalized cross-correlation matching and a relaxation method at corner points detected with the Harris-Stephens corner detector with a matching constraint based on epipolar geometry (Zhang, Deriche, et al. 1995). Images are assumed to be formed by a perspective projection. The epipolar constraint is the only known geometric constraint available for matching images taken from differing viewpoints. The method is designed for images taken with an uncalibrated camera. The implicit assumption, however, is that the scene is static.

The relaxation process is based on a match support measure that allows a higher tolerance to image plane rigid transformation and a smaller contribution from distant matches than closer ones. The match support measure determines if a candidate match has a sufficient number of neighbouring matches in both images. Since mismatches do exist, the matching process must be iterated, i.e., a relaxation process is required until all the detectable mismatches are eliminated and match neighbourhoods are stable. The update strategy only selects those points that have both high matching support and low ambiguity. The strategy is different from classical winner-take-all that easily falls into local minima and loser-take-nothing that usually converges very slowly.

Given the initial set of candidate matches, subsets of the matches are used to compute the fundamental matrix which encodes the position of the epipoles and the homography that maps the pencil of epipolar lines from one image to the next. This process is repeated using random subsets of the candidate matches in a least median of squares (LMedS) framework. The final result from the LMedS estimator is the fundamental matrix with outlier matches eliminated. The least median of squares estimator can tolerate up to 50 percent outliers.

Using the estimated epipolar geometry, further search restricted to epipolar lines (thus an efficient 1D search) is made for more unambiguous matches in a stereo matching step. Given the set of computed matches the nonreference image is resampled into the image coordinate frame of the reference image as described above.

3. EXPERIMENTAL RESULTS

The image stabilization task can be specified in general terms as follows. Given an image sequence usually consisting of at least 5 to 10 seconds of data nominally at 30 frames per second, a special frame called the reference frame is chosen at or near the beginning of the sequence. Frames subsequent to the reference frame shall be registered to the reference frame in such a way that the frames are precisely aligned with the reference image.

Figure 1 shows frames 01 and 20 from the sequence DIM01. Note the large-scale change due to the camera translation especially in the foreground and a non-trivial amount of camera rotation.

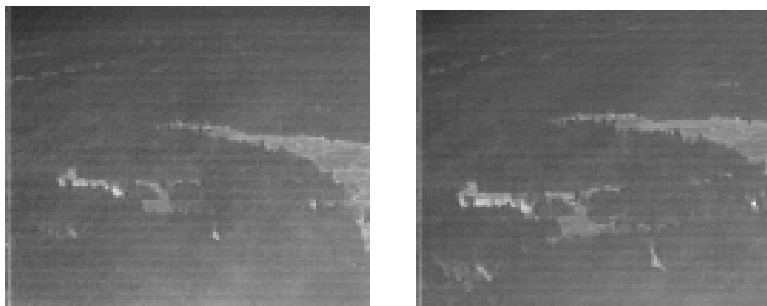


Figure 1: Left: Frame 01. Right: Frame 20 from sequence DIM01.

Figure 2 gives the GDI matching results for frames 01 and 20. The scale of the smoothing kernel had to be increased to 6 pixels from 3 pixels over the series of 20 frames mainly to compensate for the lack of fidelity of the actual image transformation with the assumed transformation. The GDI model assumes the signal is transformed by a similarity transformation. The actual model is better approximated by a local affine transformation that in turn is a local approximation

of the effect of the projective distortion due to the translating camera. Also, because the scale of the signal is increasing, more detail is present in the nonreference image that leads to a perturbation of the invariants. Additional smoothing of the reference and nonreference images reduces the perturbing effects of the scale change.

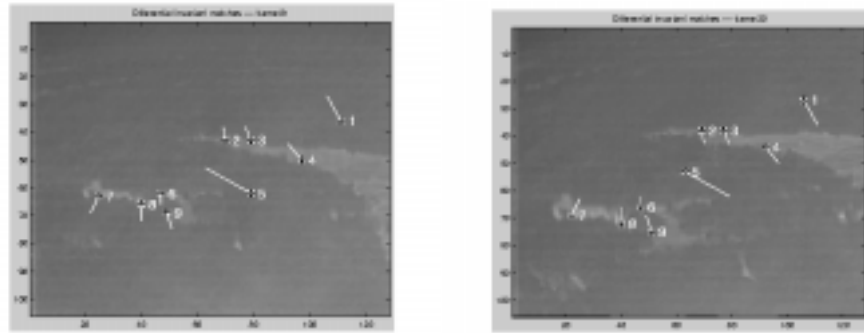


Figure 2: Greylevel differential invariant matches. Matches 4, 5 are incorrect and match 1 is mislocalized by ~3 pixels.

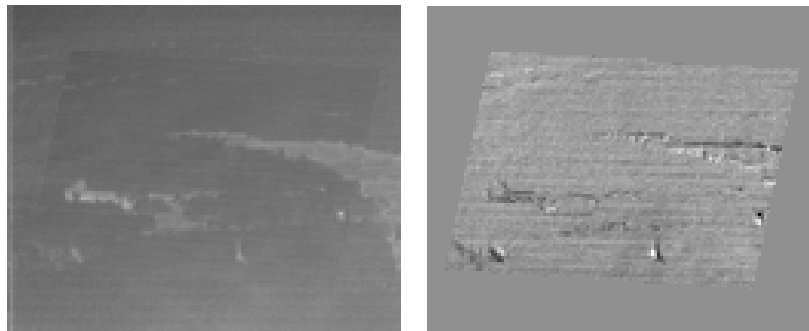


Figure 3: Left: Registered frame 20, compare with frame 01 in Fig. 1. Right: Difference of reference and registered frames.

The left of Figure 3 shows frame 20 resampled into the image coordinate system of frame 01 in accordance with the global affine transformation estimated from the four correct matches 6, 7, 8, and 9 from Figure 2. These matches are automatically returned by a relaxation process that removes outliers. Resampling was by bilinear interpolation. We see that the resampled frame 20 is close to the reference frame 01. The right of Figure 2 is the difference between the registered frame 20 and the reference frame 01. The contrast of the image is stretched to fill the range 0 to 255. The widths of brightness discontinuity highlights the error in the registration. The errors are mostly due to the clustering of the matches towards the lower left side of the image.

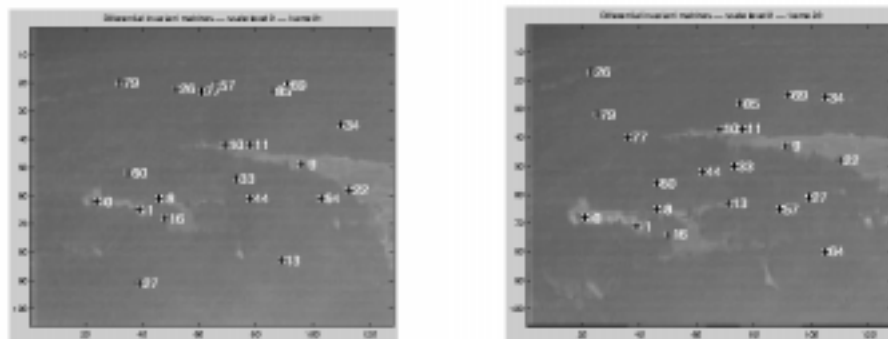


Figure 4: GDI matches for one of three scales. Matches not found in Fig. 2 were not matched over the three scale levels.

Figure 4 shows the matches returned by the GDI matching method for one of the three scale levels with the derivative of Gaussian filter scale equal to 6 pixels. The remaining two reference scale levels are 6.6 and 7.26 pixels respectively. We note the larger number of matches, 21 in all. Of these 21 matches 8 are correct, a ratio of 38%. Match 33 is correct but does not appear at all three scales. For the 6.6-pixel scale level the number of matches is 18 of which 8 are correct (44%), and for the last scale level there are 9 out of 21 matches correct (43%). With scale-space tracking there are 7

out of 9 correct matches (78%) a significant improvement in terms of the ratio of correct to total matches. Finally, the average number of records searched over both k -d trees over the three reference scale levels with an average of 860 records in each tree was 68 in keeping with the expected logarithmic time to complete a query.

The method due to Zhang *et al.* was also applied to the DIM01 sequence using the executable program provided freely via file transfer protocol by Zhang. It has been assumed that normalized cross-correlation will fail for matching images with large changes in scale and with image rotation. It was necessary to adjust the correlation parameters significantly in order to increase the methods tolerance. The correlation threshold for valid matches had to be reduced to 0.55 from 0.8, the default value. Similarly, the relaxation neighbourhood distortion factor had to be increased. The method produces only a small set of matches, however, the number of correct matches is large due to the geometric verification process applied to the correlation-based matches. Among ten matches two of them are incorrect. The geometric verification process did not eliminate them because their displacement is parallel to the local epipolar line geometry estimated from all the matches.

The mismatches were manually edited from the match list and frame 20 was registered. There is a significant misregistration towards the bottom of the image, despite the presence of several matches in that area. It is most likely caused by the projective distortion components of the image transformation, which is not properly modeled by the fitted global affine transformation.

4. CONCLUSION

We have shown two point-based methods for image matching and applied them to the stabilization of an infrared image sequence. In the GDI method the scale-space filtering enhancements produces a high proportion of correct matches, and the use of k -d tree nearest neighbour searching reduces the search time to the expected logarithmic complexity. Both methods produced good stabilization for most image sequences. A multi-scale analysis should lead to a much larger number of correct matches as stable features are detected and matched through scale-space.

When the infrared image sequence has low signal to noise ratio, lack of scene structure, and significant projective distortion, the number of correct matches is typically too sparse and not well distributed throughout the field of view. This is the case for sequence DIM01, which contains large homogenous regions and significant viewpoint changes. A solution would be to incorporate a dense matching method such as optical flow. A large change in viewpoint can lead to an image transformation that can only be modeled accurately by a 3D projective transformation instead of a 2-D global affine transformation. The method of Zhang *et al.* yields approximately the same number and quality of matches as the GDI method. However, much effort must be expended to determine the optimal thresholds for the matching metric, e.g., correlation thresholds and neighbourhood geometric distortion factors.

5. ACKNOWLEDGMENTS

Funding for this research is provided by the Natural Sciences and Engineering Research Council of Canada, Department of National Defense, and Lockheed Martin Electronic Systems Canada.

6. REFERENCES

- (Harris and Stephens 1988) Harris, C., Stephens, M., A combined corner and edge detector, Proceedings 4th Alvey Vision Conference, 147-151, 1988.
- (Horn 1986) Horn, B.K.P., Robot Vision, Cambridge, MA, The MIT Press, 1986.
- (Maheux 1998) Maheux, J., Video-rate image stabilization, Opto-Contact 98, Quebec, Quebec, July 1998.
- (Marr 1982) Marr, D., Vision: a computational investigation into the human representation and processing of visual information, Freeman, San Francisco, CA, 1992.
- (McReynolds and Lowe 1996) McReynolds, D.P., Lowe, D.G., Rigidity checking of 3D point correspondences under perspective projection, IEEE Transactions PAMI, (18)12, 1174-1185, 1996.
- (McReynolds 1997) McReynolds, D.P., Rigidity checking for matching 3D point correspondences under perspective projection, Ph.D. dissertation, University of British Columbia, 1997.
- (Schmid and Mohr 1995) Schmid, C., Mohr, R., Matching by local invariants, Rapport de Recherche, N 2644, INRIA, 1995.
- (Sproull 1991) Sproull, R.F., Refinements to nearest-neighbour searching in k -dimensional trees, Algorithmica, 6, pp. 579-589, 1991.
- (Weng *et al.* 1992) Weng, J., Ahuja, N., Huang, T.S., Matching two perspective views, IEEE Trans. PAMI, Vol. 14, No. 8, 806-825, 1992.
- (Zhang, Deriche *et al.* 1995) Zhang, Z., Deriche, R., Faugeras, O., Quang-Tuan, L., A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry, Artificial Intelligence, 78(1-2), 87-119, 1995.